Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Heterogeneous Metric Learning with Joint Graph Regularization for Cross-Media Retrieval

Xiaohua Zhai and **Yuxin Peng*** and **Jianguo Xiao**
Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
{*zhaixiaohua, pengyuxin, xiaojianguo*}*@pku.edu.cn*

Abstract

As the major component of big data, unstructured heterogeneous multimedia content such as text, image, audio, video and 3D increasing rapidly on the Internet. User demand a new type of cross-media retrieval where user can search results across various media by submitting query of any media. Since the query and the retrieved results can be of different media, how to learn a *heterogeneous metric* is the key challenge. Most existing metric learning algorithms only focus on a single media where all of the media objects share the same data representation. In this paper, we propose a joint graph regularized heterogeneous metric learning (JGRHML) algorithm, which integrates the structure of different media into a joint graph regularization. In JGRHML, different media are complementary to each other and optimizing them simultaneously can make the solution smoother for both media and further improve the accuracy of the final metric. Based on the heterogeneous metric, we further learn a high-level semantic metric through label propagation. JGRHML is effective to explore the semantic relationship hidden across different modalities. The experimental results on two datasets with up to five media types show the effectiveness of our proposed approach.

Introduction

As the major component of big data, unstructured heterogeneous multimedia content such as text, image, audio, video and 3D increasing rapidly on the Internet. Many research efforts have been devoted to the content-based multimedia retrieval (Jeon, Lavrenko, and Manmatha 2003; Greenspan, Goldberger, and Mayer 2004; Escalante et al. 2008). However, the prevailing methods are single-media retrieval and multi-modal retrieval. For the former one, the retrieved result and user query are of the same media, such as text retrieval, image retrieval, audio retrieval and video retrieval. For the latter one, the retrieved result and user query share the same multiple media, which are combined together to achieve better result, such as using image and text to retrieve image and text results. In summary, existing methods generally focus on the single-media relationship but ignore the cross-media relationship between different modalities, which is important for better understanding the multimedia content. What's more, these methods cannot support the content-based cross-media retrieval, such as using image to retrieve relevant text, audio, video and 3D.

In fact, users demand such cross-media retrieval to search results across various modalities by submitting query of any media type. On one hand, user can obtain all of the related results at one time, which is more comprehensive than traditional retrieval methods. Suppose we are on a visit to the Golden Gate Bridge, by taking a photo, cross-media retrieval is able to retrieve all of the textual materials, audio commentary and visual guides for us. It will help us get familiar with the Golden Gate Bridge quickly. On the other hand, users can submit any media content at hand as query, which is very convenient. For example, we can obtain the singing sound of an unfamiliar bird just by a textual description, without recording of a similar sound. Content-based cross-media retrieval is an interesting, yet difficult question. The similarity measure between homogeneous media objects has always been a difficult problem. So how to measure the content similarity between heterogeneous media objects is much more challenge. In this paper, we focus on learning a *heterogeneous metric* between heterogeneous media objects.

Recently, learning from heterogeneous data has shown effectiveness in heterogeneous transfer learning(Zhu et al. 2011), heterogeneous multi task learning(Zhang and Yeung 2011), and transductive classification on heterogeneous information networks(Ji et al. 2010). However, an underlying assumption shared by most of the metric learning methods (Bar-Hillel et al. 2005; Davis et al. 2007; K.Weinberger, Blitzer, and Saul 2006; Xing et al. 2002; Hoi, Liu, and Chang 2008; Park et al. 2011) is that all of the data share the same data representation. So existing metric learning methods have previously been designed primarily for single-media data and cannot be directly applied to cross-media data. The similarity relation between heterogeneous media objects is not a metric, hence, does not fall into the standard framework of metric learning (Bronstein et al. 2010). Little attention has been paid to heterogeneous metric learning for cross-media data analysis in the literature. The reason may be that heterogeneous metric learning is more challenging than homogeneous metric learning because of the requirement to mining the similarity between heteroge-

*Corresponding author.

neous media objects. A possible solution is to linearly map the heterogeneous objects into a new space. Then the dot product in the new space is defined as the similarity function(Wu, Xu, and Li 2010). However, only the matched and mismatched pairs are explored. They cannot make full use of the structure information of the whole heterogeneous spaces.

In this paper, we propose a joint graph regularized heterogeneous metric learning (JGRHML) algorithm, which integrates the structure of different media into a joint graph regularization to better exploit the structure information. More importantly, the joint graph regularization makes the learned data projection of each media type consistent with the original graph Laplacian. Different media are complementary to each other in the joint graph regularization and optimizing them simultaneously can make the solution smoother for both media. In addition, we further learn an explicit high-level semantic representation through label propagation based on a unified k-nearest neighbor graph, which is constructed from all of the labeled and unlabeled heterogeneous data. Therefore, both heterogeneous similarities and homogeneous similarities are incorporated into the unified graph, which can explore the cross-media correlation among all of the media objects of different media types. To the best of our knowledge, our method has made the first attempt to heterogeneous metric learning with joint graph regularization. Experiments on two datasets with up to five media types show the effectiveness of our proposed approach, as compared with the state-of-the-art methods.

The rest of this paper will be organized as follows. In Section 2, we demonstrate the problem definition and preliminaries. In section 3, we introduce the joint graph regularized heterogeneous metric. Section 4 shows the experimental results. Finally, we conclude this paper in Section 5.

Problem Definition and Preliminaries

In this section, we first define the problem to be addressed. Then we briefly review the metric learning for homogeneous data.

Definition 1 *(Content-based Cross-Media Retrieval) Given a dataset with heterogeneous multimedia content* $\mathbb{D} = \{(x_1, l_1^x), ..., (x_m, l_m^x), (y_1, l_1^y), ..., (y_n, l_n^y)\}$ *consists of* $m + n$ *media objects. Here* $x_i \in \mathcal{X}$ *and* $y_i \in \mathcal{Y}$ *denote different media types* $\mathcal{X}$ *and* $\mathcal{Y}$*, which can be image, text, audio, video or 3D.* x_i *and* y_i *are labeled as* l_i^x *and* l_i^y*. The basic goal of content-based cross-media retrieval is to retrieve relevant x in unlabeled dataset* $\mathbb{T} = \{x_1, ..., x_p, y_1, ..., y_q\}$ *in response to the query of y and vice-versa.*

For example, a set of text articles can be regarded as $\mathcal{X}$, and a set of images can be regarded as $\mathcal{Y}$. The task is to retrieve text articles in response to the query of images and vice-versa. Note that once the content-based cross-media retrieval problem is solved, we can retrieve all of the related heterogeneous multimedia content by submitting query of any media type.

A Brief Review of Metric Learning We first review the framework of traditional homogeneous metric learning methods. Given two homogeneous media objects x_i and x_j, it aims to learn an optimal metric $d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{A}(x_i - x_j)}$ according to the similarity constraints and dissimilarity constraints. It assumes that there is some corresponding linear transformation U for a possible metric $d_A(x_i, x_j)$(Hoi, Liu, and Chang 2008). As a result, the distance between two input examples can be computed as follows:

$$d(x_i, x_j) = \sqrt{(\mathbf{U}^T x_i - \mathbf{U}^T x_j)^T (\mathbf{U}^T x_i - \mathbf{U}^T x_j)} \quad (1)$$

The goal of metric learning is to learn the linear transformation $\mathbf{U}$ for the original features.

Joint Graph Regularized Heterogeneous Metric

Heterogeneous Metric Learning

We can obtain two sets of heterogeneous pairwise constraints among the heterogeneous media objects:

$$\begin{aligned} \mathcal{S} &= \{(x_i, y_j) | l_i^x = l_j^y\} \\ \mathcal{D} &= \{(x_i, y_j) | l_i^x \neq l_j^y\} \end{aligned} \quad (2)$$

where $\mathcal{S}$ is the set of similarity constraints, and $\mathcal{D}$ is the set of dissimilarity constraints. Each pairwise constraint (x_i, y_j) indicates if two heterogeneous media objects x_i and y_j are relevant or irrelevant inferred from the category label. We denote both $\mathcal{S}$ and $\mathcal{D}$ on the dataset $\mathbb{D}$ with a single matrix $\mathbf{Z} = \{z_{ij}\}_{m \times n}$:

$$z_{ij} = \begin{cases} 1, & (x_i, y_j) \in \mathcal{S}; \\ -1, & (x_i, y_j) \in \mathcal{D}. \end{cases} \quad (3)$$

For any two given heterogeneous media objects x_i and y_j ,let $d(x_i, y_j)$ denote the heterogeneous distance between them. The similarity relation between heterogeneous data is not a metric, hence, does not fall into the standard framework of metric learning (Bronstein et al. 2010). To learn the cross-media similarity metric, we propose a new learning approach to handle such heterogeneous similarity problems. Instead of learning a single transformation $\mathbf{U}$ of the input, we propose to learn multiple linear transformation matrices $\mathbf{U}$ and $\mathbf{V}$, which map to the same output space. Let $\mathbf{U} \in \mathbb{R}^{d^x * c}$, $\mathbf{V} \in \mathbb{R}^{d^y * c}$ be the distance parameter matrices for $\mathbf{X} \in \mathbb{R}^{d^x * m}$ and $\mathbf{Y} \in \mathbb{R}^{d^y * n}$ respectively, here d^x, d^y are the dimensions of original media types, c is the dimension of mapped space, m and n are the number of media objects of media $\mathcal{X}$ and media $\mathcal{Y}$ respectively. We define the proposed heterogeneous distance measure as follows:

$$d(x_i, y_j) = \sqrt{(\mathbf{U}^T x_i - \mathbf{V}^T y_j)^T (\mathbf{U}^T x_i - \mathbf{V}^T y_j)} \quad (4)$$

We aim to learn two parameter matrices $\mathbf{U}_{d^x * c}, \mathbf{V}_{d^y * c}$ from the training heterogeneous multimedia dataset $\{\mathbf{X}_{d^x * m}, \mathbf{Y}_{d^y * n}\}$. Frequently used notations and descriptions are summarized in Table 1.

Objective Function

We formulate a general regularization framework for heterogeneous distance metric learning as follows:

$$\underset{\mathbf{U},\mathbf{V}}{\operatorname{argmin}} \; f(\mathbf{U}, \mathbf{V}) + \omega g(\mathbf{U}, \mathbf{V}) + \lambda r(\mathbf{U}, \mathbf{V}) \quad (5)$$

Table 1: Notations and descriptions used in this paper.

Notation	Description
m, n	#training examples of media object x, y
p, q	#testing examples of media object x, y
d^x	feature dimension of media object x
d^y	feature dimension of media object y
λ, ω	regularization parameters
$\mathbf{X}$	$d^x \times m$ data matrix of media object x
$\mathbf{Y}$	$d^y \times n$ data matrix of media object y
$\mathbf{Z}$	$m \times n$ matrix with heterogeneous constrains
$\mathbf{U}$	$d^x \times c$ transformation matrix for media x
$\mathbf{V}$	$d^y \times c$ transformation matrix for media y
$\mathbf{O}$	$c \times (m+n)$ data matrix for all media objects

where $f(\mathbf{U}, \mathbf{V})$ is the loss function defined on the sets of similarity and dissimilarity constraints $\mathcal{S}$ and $\mathcal{D}$, $g(\mathbf{U}, \mathbf{V})$ and $r(\mathbf{U}, \mathbf{V})$ are regularizer defined on the target parameter matrices $\mathbf{U}, \mathbf{V}$. $\lambda > 0, \omega > 0$ are the balancing parameters.

Loss function The loss function $f(\mathbf{U}, \mathbf{V})$ should be defined in the way such that the minimization of the loss function will result in minimizing (maximizing) the distances between the media objects with the similarity (dissimilarity) constraints. In this paper, we adopt the sum of squared distances expression for defining the loss functions in terms of its effectiveness and efficiency in practice:

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ij} \|\mathbf{U}^T x_i - \mathbf{V}^T y_j\|^2 \tag{6}$$

where z_{ij} is defined in equation (3). To balance the influence of similarity constraints and dissimilarity constraints, we normalize the elements of $\mathbf{Z}$ column by column to make sure that the sum of each column is zero.

Scale regularization We define the regularization item $r(\mathbf{U}, \mathbf{V})$ as follows:

$$r(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{U}\|_F^2 + \frac{1}{2} \|\mathbf{V}\|_F^2 \tag{7}$$

where $\|\mathbf{U}\|_F^2$ and $\|\mathbf{V}\|_F^2$ are used to control the scale of the parameter matrices and reduce overfitting.

Joint graph regularization Next, we will introduce the joint graph regularization item $g(\mathbf{U}, \mathbf{V})$. We found that the similarity constraints in both modalities are helpful for metric learning. So we try to make the learned transformation consistent with the similarity constraints in both modalities.

For heterogeneous data with multiple representations, we define a joint undirected graph, $G = (V, \mathbf{W})$ on the dataset. Each element w_{ij} of the similarity matrix $\mathbf{W} = \{w_{ij}\}_{(m+n)*(m+n)}$ means the similarity between the i-th media object and j-th media object. Note that all of the heterogeneous media objects $o_i \in \mathbb{D}, i = 1, ..., m+n$ are incorporated into the joint graph. Here, we adopt the label information to construct the symmetric similarity matrix:

$$w_{ij} = \begin{cases} 1, & l_i == l_j \wedge i \neq j; \\ 0, & otherwise. \end{cases} \tag{8}$$

Where $l_i = l_i^x$ for $0 < i \leq m$ and $l_i = l_{(i-m)}^y$ for $m < i \leq m+n$. We set $w_{ii} = 0$ for $1 \leq i \leq m+n$ to avoid self-reinforcement. The normalized graph Laplacian L is defined as:

$$\bar{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2} \tag{9}$$

where $\mathbf{I}$ is an $(m+n) \times (m+n)$ identity matrix and $\mathbf{D}$ is an $(m+n) \times (m+n)$ diagonal matrix with $d_{ii} = \sum_j w_{ij}$. It should be noted that $\bar{\mathbf{L}}$ is symmetric and positive semidefinite, with eigenvalues in the interval $[0, 2]$(Chung 1997). We define:

$$\begin{aligned} \mathbf{O} &= \begin{pmatrix} \mathbf{U}^T\mathbf{X} & \mathbf{V}^T\mathbf{Y} \end{pmatrix} \\ \bar{\mathbf{L}} &= \begin{pmatrix} \bar{\mathbf{L}}^x & \bar{\mathbf{L}}^{xy} \\ \bar{\mathbf{L}}^{yx} & \bar{\mathbf{L}}^y \end{pmatrix} \end{aligned} \tag{10}$$

where $\mathbf{O}$ represents for all of the media objects in the learned metric space, $\bar{\mathbf{L}}$ denotes the normalized graph Laplacian. Based on this, we formulate the regularization as follows:

$$\begin{aligned} g(\mathbf{U}, \mathbf{V}) &= \frac{1}{4} \sum_{i,j=1}^{m+n} \| \frac{o_i}{\sqrt{d_{ii}}} - \frac{o_j}{\sqrt{d_{jj}}} \|^2 w_{ij} \\ &= \frac{1}{2} tr(\mathbf{O}\bar{\mathbf{L}}\mathbf{O}^T) \\ &= \frac{1}{2} tr(\mathbf{U}^T\mathbf{X}\bar{\mathbf{L}}^x\mathbf{X}^T\mathbf{U}) + \frac{1}{2} tr(\mathbf{U}^T\mathbf{X}\bar{\mathbf{L}}^{xy}\mathbf{Y}^T\mathbf{V}) \\ &+ \frac{1}{2} tr(\mathbf{V}^T\mathbf{Y}\bar{\mathbf{L}}^{yx}\mathbf{X}^T\mathbf{U}) + \frac{1}{2} tr(\mathbf{V}^T\mathbf{Y}\bar{\mathbf{L}}^{y}\mathbf{Y}^T\mathbf{V}) \end{aligned} \tag{11}$$

where $tr(\mathbf{X})$ is the trace of a matrix $\mathbf{X}$. The regularization $g(\mathbf{U}, \mathbf{V})$ penalizes large changes of the mapping function $\mathbf{U}, \mathbf{V}$ between two nodes linked with a large weight. In other words, minimizing $g(\mathbf{U}, \mathbf{V})$ encourages the smoothness of a mapping over the joint data graph, which is constructed from the initial label information.

Iterative optimization

We propose an iterative method to minimize the above objective function (5). Firstly, we initialize $\mathbf{U}$ and $\mathbf{V}$ by cross-media factor analysis(Li et al. 2003). We simply assume that the media objects with the same label should have similar representations. It finds the optimal transformations that can best represent the coupled patterns between features of two different subsets. We want orthogonal transformation matrices $\mathbf{U}$ and $\mathbf{V}$ that can minimize the following object function:

$$\begin{aligned} &\|\mathbf{U}^T\mathbf{X}' - \mathbf{V}^T\mathbf{Y}'\|^2 \\ &s.t. \quad \mathbf{U}\mathbf{U}^T = \mathbf{I}, \mathbf{V}\mathbf{V}^T = \mathbf{I}. \end{aligned} \tag{12}$$

where $\mathbf{X}'$ and $\mathbf{Y}'$ represent for two sets of coupled media objects from different media with the same labels. $\mathbf{U}$ and $\mathbf{V}$ define two orthogonal transformation spaces where media objects in $\mathbf{X}'$ and $\mathbf{Y}'$ can be projected as close to each other as possible. We have:

$$\begin{aligned} &\|\mathbf{U}^T\mathbf{X}' - \mathbf{V}^T\mathbf{Y}'\|^2 = \\ &tr(\mathbf{X}'^T\mathbf{X}') + tr(\mathbf{Y}'^T\mathbf{Y}') - 2tr(\mathbf{X}'^T\mathbf{U}\mathbf{V}^T\mathbf{Y}') \end{aligned} \tag{13}$$

where $tr(\mathbf{X})$ is the trace of a matrix $\mathbf{X}$. We can easily see from above that matrices $\mathbf{U}$ and $\mathbf{V}$ which maximize

$tr(\mathbf{X}'^T\mathbf{U}\mathbf{V}^T\mathbf{Y}')$ will minimize (13). Such matrices are given by singular value decomposition:

$$\mathbf{X}'\mathbf{Y}'^T = \mathbf{U}\Sigma\mathbf{V} \tag{14}$$

Once the initial value of $\mathbf{U}$ and $\mathbf{V}$ are given, in each iteration, we first update $\mathbf{U}$ given $\mathbf{V}$ and then update $\mathbf{V}$ given $\mathbf{U}$. These two alternating steps are described as below.

Fix V and update U. Let $Q(\mathbf{U},\mathbf{V})$ denote the objective function in equation (5). Differentiating $Q(\mathbf{U},\mathbf{V})$ with respect to $\mathbf{U}$ and setting it to zero, we have the following equation:

$$\begin{aligned}\frac{\partial Q(\mathbf{U},\mathbf{V})}{\partial \mathbf{U}} = &\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}(x_i x_i^T\mathbf{U} - x_i y_j^T\mathbf{V}) + \\ &\omega\mathbf{X}\bar{\mathbf{L}}^x\mathbf{X}^T\mathbf{U} + \omega\mathbf{X}\bar{\mathbf{L}}^{xy}\mathbf{Y}^T\mathbf{V} + \lambda\mathbf{U} = 0\end{aligned} \tag{15}$$

which can be transformed into:

$$\begin{aligned}&(\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}x_i x_i^T + \omega\mathbf{X}\bar{\mathbf{L}}^x\mathbf{X}^T + \lambda\mathbf{I})\mathbf{U} \\ &= (\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}x_i y_j^T - \omega\mathbf{X}\bar{\mathbf{L}}^{xy}\mathbf{Y}^T)\mathbf{V}\end{aligned} \tag{16}$$

we could obtain the analytical solution as follows:

$$\begin{aligned}\mathbf{U} =&(\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}x_i x_i^T + \omega\mathbf{X}\bar{\mathbf{L}}^x\mathbf{X}^T + \lambda\mathbf{I})^{-1} \\ &(\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}x_i y_j^T - \omega\mathbf{X}\bar{\mathbf{L}}^{xy}\mathbf{Y}^T)\mathbf{V}\end{aligned} \tag{17}$$

Fix U and update V. Similarly, differentiating $Q(\mathbf{U},\mathbf{V})$ with respect to $\mathbf{V}$ and setting it to zero, we could obtain the analytical solution:

$$\begin{aligned}\mathbf{V} =&(\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}y_j y_j^T + \omega\mathbf{Y}\bar{\mathbf{L}}^y\mathbf{Y}^T + \lambda\mathbf{I})^{-1} \\ &(\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij}y_j x_i^T - \omega\mathbf{Y}\bar{\mathbf{L}}^{yx}\mathbf{X}^T)\mathbf{U}\end{aligned} \tag{18}$$

We alternate between updates to $\mathbf{U}$ and $\mathbf{V}$ for several iterations to find a locally optimal solution. Here the iteration continues until the cross-validation performance decreases on the training set. In practice, the iteration only repeats several rounds.

High-Level Semantic Metric

So far, we have obtained the parameter matrices $\mathbf{U}$ and $\mathbf{V}$. In this section, we further learn an explicit high-level semantic (label probabilities) representation through label propagation based on a unified k-NN graph, which is constructed from both heterogeneous similarities and homogeneous similarities. The final semantic space is represented as $\mathbb{R}^s$, where s is the number of categories. Each dimension of $o_i \in \mathbb{R}^s$ is regarded as the probability of a media object belonging to the corresponding semantic category. It can explore the semantic relationship hidden across different modalities, which further improves the performance of our proposed heterogeneous metric.

Algorithm 1 Joint Graph Regularized Heterogeneous Metric Learning

Require: Heterogeneous training set $\{\mathbf{X}_{d^x*m}, \mathbf{Y}_{d^y*n}\}$ and testing set $\{\mathbf{X}_{d^x*p}, \mathbf{Y}_{d^y*q}\}$.
Ensure: Heterogeneous similarity matrix $\mathbf{H}$ where $H(i,j) = Sim(x_i, y_j)$.
1: Initialize projection matrices $\mathbf{U}_{d^x*c}$ and $\mathbf{V}_{d^y*c}$ according to equation (14).
2: Update $\mathbf{U}$ and $\mathbf{V}$ iteratively according to the equation (17) and equation (18).
3: Project the heterogeneous feature in $\mathbb{R}^c$ according to $\mathbf{U}$ and $\mathbf{V}$.
4: Construct unified k-NN graph accordign to equation (19).
5: Further learn the explicit high-level semantic representation in $\mathbb{R}^s$ according to the equation (20).
6: Obtain the similarity between media objects according to equation (21).

Recall that we are given a labeled dataset $\mathbb{D} = \{o_1, ...o_{m+n}\}$ and an unlabel dataset $\mathbb{T} = \{o_1, ...o_{p+q}\}$. Here we use $o_i \in \{\mathcal{X}, \mathcal{Y}\}$ to represent all of the labeled and unlabeled heterogeneous media objects. The goal is to map the media objects $o_i \in \mathbb{R}^c$ into the high-level semantic space $\mathbb{R}^s$ according to the similarities between all of the labeled and unlabeled heterogeneous media objects.

Let $\mathcal{F}$ denote the set of $(m+n+p+q)\times s$ matrices, here $m+n+p+q$ stands for the number of the media objects of all of the heterogeneous media objects, s represents the number of categories. Define a $(m+n+p+q)\times s$ matrix $\mathbf{Y} \in \mathcal{F}$ where $\mathbf{Y}_{ij} = 1$ if o_i is labeled as $l_i = j$ and $\mathbf{Y}_{ij} = -1$ if $l_i \neq j$ for labeled data ($i = 1, 2, ...m+n$), $\mathbf{Y}_{ij} = 0$ for unlabeled data ($i = m+n+1, ...m+n+p+q$). Clearly, $\mathbf{Y}$ is consistent with the initial training labels according the decision rule. The algorithm for learning the high-level semantic through label propagation using a unified k-NN graph is as follows:

(1) Form the affinity matrix $\mathbf{W}$ of the unified k-NN graph $w_{ij} = \sigma(K(o_i, o_j))$ if $o_j (j \neq i)$ is among the k-nearest neighbors of o_i and $w_{ij} = 0$ otherwise. $\sigma(z) = (1 + exp(-z))^{-1}$ is the sigmoid function, $K(o_i, o_j)$ is the similarity function between two media objects, which is defined as follows:

$$\begin{cases} -\|\mathbf{U}^T o_i - \mathbf{U}^T o_j\|, & \{o_i, o_j\} \subseteq \mathcal{X}; \\ -\|\mathbf{V}^T o_i - \mathbf{V}^T o_j\|, & \{o_i, o_j\} \subseteq \mathcal{Y}; \\ -\|\mathbf{U}^T o_i - \mathbf{V}^T o_j\|, & o_i \in \mathcal{X} \wedge o_j \in \mathcal{Y}; \\ -\|\mathbf{V}^T o_i - \mathbf{U}^T o_j\|, & o_i \in \mathcal{Y} \wedge o_j \in \mathcal{X}. \end{cases} \tag{19}$$

(2) Construct the matrix $\bar{\mathbf{S}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of $\mathbf{W}$.

(3) Iterate $\mathbf{F}(m+1) = \alpha\bar{\mathbf{S}}\mathbf{F}(m) + (1-\alpha)\mathbf{Y}$ until convergence, where $\mathbf{F}(m)$ denotes the propagation result and we set $F(0) = \mathbf{Y}$, α is a parameter in the range (0,1). Here we set $\alpha = 0.1$ empirically and normalize the elements of

$\mathbf{Y}$ column by column to make sure that the sum of each column is zero.

(4) Let $\mathbf{F}^*$ denote the limit of the sequence $\{\mathbf{F}(m)\}$, which represents the high-level semantic representation.

According to (Zhou et al. 2003), the above algorithm converges to:

$$\mathbf{F}^{(s)^*} = (1-\alpha)(\mathbf{I}-\alpha\mathbf{W})^{-1}\mathbf{Y} \tag{20}$$

Once the high-level representation is obtained, for $\{o_i, o_j\} \subseteq \mathbb{R}^s$, the similarity measure is defined as follows:

$$Sim(o_i, o_j) = o_i \cdot o_j = \sum_{l \in \mathcal{L}} o_i(l) o_j(l) \tag{21}$$

where $\cdot$ denotes the element-wise multiplication. Since l-th dimension of o_i represents for the probability of o_i belonging to category l, equation (21) actually measures the probability of two media objects belonging to the same semantic category.

In summary, the proposed JGRHML algorithm is listed in Algorithm 1.

Experiments

In this section, we conduct experiments on two real-world datasets to verify the effectiveness of our proposed JGRHML method.

Datasets

Cross-media retrieval is a relatively new problem. There are few publicly available cross-media datasets. A notable publicly available cross-media dataset is Wikipedia dataset (Rasiwasia et al. 2010). However, it only includes images and texts, which cannot fully evaluate the retrieval performance on multiple media types, such as using image to retrieve relevant text, image, audio, video and 3D. To evaluate the performance objectively, we further construct a new XMedia dataset, which contains up to five media types, i.e., text, image, audio, video and 3D. To the best of our knowledge, XMedia dataset is the first cross-media dataset consists of five media types. Because "X" looks like cross line, XMedia stands for cross-media retrieval among all the different media types. Following we will introduce the above two datasets in detail.

Wikipedia dataset (Rasiwasia et al. 2010) is chosen from the Wikipedia's "featured articles". This is a continually updated collection of 2700 articles that have been selected and reviewed by Wikipedia's editors since 2009. Each article is accompanied with one or more images from Wikimedia Commons. Each article is split into several sections according to its section headings.The dataset contains a total of 2866 documents, which are text-image pairs and annotated with a label from the vocabulary of 10 semantic categories. The dataset is randomly split into a training set of 2173 documents and a test set of 693 documents.

XMedia dataset consists of 5000 texts, 5000 images, 1000 audio, 500 videos and 500 3D models. All of the media objects are crawled from the Internet. In detail, all of the texts are crawled from the Wikipedia articles and the videos are crawled from the Youtube website. The other media objects consist of two parts: 800 images from Wikipedia articles and 4200 images from the photo sharing website Flickr, 800 clips of audio from freesound website and 200 clips of audio from findsound website, 3D models are crawled from 3D Warehouse website and Princeton 3D Model Search Engine website. This dataset is organized into 20 categories with 600 media objects per category. The dataset is randomly split into a training set of 9600 media objects and a test set of 2400 media objects.

For the two datasets, bag-of-words (BOW) model and topic model are utilized to represent the images and text respectively. Each image is represented using a histogram of a 128-codeword SIFT codebook and each text is represented using a histogram of a 10-topic latent Dirichlet allocation (LDA) model, exactly the same as (Rasiwasia et al. 2010). In addition, we adopt 29-dim MFCC features to represent each clip of audio. We segment each clip of video into video shots. Then 128-dimension BoW histogram features are extracted for each video keyframe. The final similarity for video is obtained by averaging all of the similarities of the video keyframes. Each 3D model is firstly represented as the concatenated 4700-dimension vector of a set of Light-Field descriptors as described in (Chen et al. 2003). Then the concatenated vector is reduced to 128-dimension vector based on Principal Component Analysis (PCA). All of the compared methods in the experiment section adopt the same features and training data for fair comparison.

Baseline Methods and Evaluation Metrics

To evaluate objectively our proposed method, 22 cross-media retrieval tasks are conducted. In these tasks, each media object is served as the query, and the result is the ranking of heterogeneous media objects. We evaluate all of the possible combination of heterogeneous media content on Wikipedia and XMedia dataset. Four different baseline methods are compared, which are summarized as follows:

- **Random** Randomly retrieving the results.
- **CCA** Canonical correlation analysis (Hotelling 1936) is widely used for cross-media relationship analysis (Kidron, Schechner, and Elad 2005; Bredin and Chollet 2007; Blaschko and Lampert 2007; Rasiwasia et al. 2010). Through CCA we could learn the subspace that maximizes the correlation between two sets of heterogeneous data, which is a natural possible solution to analyze the correlation between two multivariate random vectors.
- **CFA** also learns a subspace for different modalities. Unlike CCA which finds transformation matrices that maximize the correlation between two subsets of features, the cross-modal factor analysis (CFA) method (Li et al. 2003) adopts a criterion of minimizing the Frobenius norm between pairwise data in the transformed domain.
- **CCA+SMN** is current state-of-the-art approach(Rasiwasia et al. 2010), since it consider not only correlation analysis but also semantic abstraction for different modalities. The correlation between different modalities is learned with CCA and abstraction is achieved by representing text and image at a more general semantic level.

Table 2: Cross-media retrieval on two datasets (MAP scores), our proposed JGRHML consistently outperforms compared methods. $\mathcal{X} \rightarrow \mathcal{Y}$ means that media $\mathcal{X}$ are served as query and results are media $\mathcal{Y}$. The upper part shows the MAP scores on Wikipedia dataset and the lower part shows the MAP scores on XMedia dataset.

Task	Random	CFA	CCA	CCA+SMN	JGRHML
Image→Text	0.118	0.246	0.249	0.277	**0.329**
Text→Image	0.118	0.195	0.196	0.226	**0.256**
Image→Text	0.057	0.127	0.119	0.141	**0.176**
Image→Audio	0.074	0.129	0.103	0.162	**0.198**
Image→Video	0.088	0.174	0.098	0.197	**0.239**
Image→3D	0.086	0.159	0.117	0.299	**0.347**
Text→Image	0.057	0.126	0.114	0.138	**0.190**
Text→Audio	0.074	0.136	0.127	0.155	**0.183**
Text→Video	0.090	0.137	0.110	0.127	**0.201**
Text→3D	0.090	0.180	0.160	0.187	**0.279**
Audio→Image	0.056	0.109	0.078	0.129	**0.177**
Audio→Text	0.057	0.110	0.106	0.122	**0.142**
Audio→Video	0.081	0.151	0.114	0.142	**0.181**
Audio→3D	0.086	0.147	0.164	0.269	**0.318**
Video→Image	0.055	0.126	0.065	0.157	**0.192**
Video→Text	0.056	0.102	0.078	0.090	**0.134**
Video→Audio	0.072	0.117	0.093	0.137	**0.139**
Video→3D	0.098	0.178	0.134	0.101	**0.253**
3D→Image	0.056	0.115	0.073	0.248	**0.296**
3D→Text	0.055	0.109	0.104	0.135	**0.183**
3D→Audio	0.066	0.147	0.153	0.214	**0.318**
3D→Video	0.097	0.186	0.123	0.101	**0.242**
Average	0.077	0.146	0.122	0.171	**0.226**

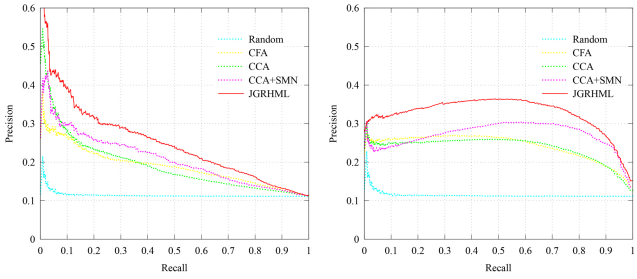
We evaluate the retrieving results with the precision-recall (PR) curves and mean average precision (MAP), which are widely used in the image retrieval literature. The MAP score is the average precision at the ranks where recall changes, and takes a value in the range [0,1]. The parameters in equation (17) and equation (18) are set according to the 5-fold cross validation on the training set. We set $\omega = 0.1, \lambda = 1000$ respectively and set $k = 90$ for the unified k-NN graph.

Experimental Results

In this section, we compare our proposed joint graph regularized heterogeneous metric learning method with four different baselines for cross-media retrieval.

Table 2 shows the MAP scores of our proposed joint graph regularized heterogeneous metric learning (JGRHML) and baseline methods on two datasets. Compared to current state-of-the-art method CCA+SMN on all of the 22 cross-media retrieval tasks, our proposed JGRHML improves the average MAP from 17.1% to 22.6%, which is inspiring. Our proposed JGRHML consistently outperforms compared methods on all of the 22 tasks, which due to the factor that JGRHML integrates the structure of different media into a joint graph regularization. So different media are complementary to each other and optimizing them simultaneously can make the solution smoother for both media and further improve the accuracy of the final metric.

The upper part of Table 2 shows the MAP scores on Wikipedia dataset. It can be seen that the MAP scores of cross-media retrieval methods all significantly outperform those of random retrieval. The performance of CFA is approaching that of CCA. CCA+SMN is current state-of-the-art method since they consider not only correlation analysis but also semantic abstraction for different media. Figure 1 shows the PR curve of the above methods. It can be seen that JGRHML also attains higher precision at most levels of recall, outperforming current state-of-the-art methods.



(a) Text Query (b) Image Query

Figure 1: Precision recall curves on Wikipedia dataset.

The lower part Table 2 shows the cross-media retrieval result on the XMedia dataset. It can be seen that our proposed JGRHML also achieves the best result so far. We find that the performance of most methods decreases on the XMedia dataset. It is reasonable because the XMedia dataset is more challenging since there are more categories and media types. So it is hard to search the result of the same semantic with the user query. However, the performance gain of our JGRHML method remained unchanged, as compared to current state-of-the-art method.

Conclusion

In this paper, we have proposed JGRHML algorithm to obtain the similarity between heterogeneous media objects. JGRHML integrates the structure of different media into a joint graph regularization, where different media are complementary to each other and optimizing them simultaneously can make the solution smoother for both media and further improve the accuracy of the final metric. Based on the heterogeneous metric, we further learn a high-level semantic metric through label propagation. In the future, on one hand, we intend to jointly modeling multiple modalities, on the other hand, we will apply heterogeneous metric learning algorithm to more applications.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61073084, Beijing Natural Science Foundation of China under Grant 4122035, Ph.D. Programs Foundation of Ministry of Education of China under Grant 20120001110097, and National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503.

References

Bar-Hillel, A.; Hertz, T.; Shental, N.; and D.Weinshall. 2005. Learning a mahanalobis metric from equivalence constraints. *JMLR* 6:937–965.

Blaschko, M., and Lampert, C. 2007. Correlational spectral clustering. *Computer Vision and Pattern Recognition (CVPR)*.

Bredin, H., and Chollet, G. 2007. Audio-visual speech synchrony measure for talking-face identity verification. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Bronstein, M.; Bronstein, A.; Michel, F.; and Paragios, N. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. 3594–3601.

Chen, D.; Tian, X.; Shen, Y.; and Ouhyoung, M. 2003. On visual similarity based 3d model retrieval. *Computer Graphics Forum* 22(3):223–232.

Chung, F. 1997. Spectral graph theory. *American Mathematical Society*.

Davis, J.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. 2007. Information-theoretic metric learning. *ICML*.

Escalante, H.; Hérnadez, C.; Sucar, L.; and Montes, M. 2008. Late fusion of heterogeneous methods for multimedia image retrieval. *Proceeding of the 1st ACM international conference on Multimedia information retrieval*.

Greenspan, H.; Goldberger, J.; and Mayer, A. 2004. Probabilistic space-time video modeling via piecewise gmm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(3):384–396.

Hoi, S.; Liu, W.; and Chang, S. 2008. Semi-supervised distance metric learning for collaborative image retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, 1–7.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3-4):321–377.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the 26th annual international ACM SIGIR conference*.

Ji, M.; Sun, Y.; Danilevsky, M.; Han, J.; and Gao, J. 2010. Graph regularized transductive classification on heterogeneous information networks. *Machine Learning and Knowledge Discovery in Databases*.

Kidron, E.; Schechner, Y.; and Elad, M. 2005. Pixels that sound. *Computer Vision and Pattern Recognition (CVPR)*.

K.Weinberger; Blitzer, J.; and Saul, L. 2006. Distance metric learning for large margin nearest neighbor classification. *NIPS*.

Li, D.; Dimitrova, N.; Li, M.; and Sethi, I. K. 2003. Multimedia content processing through cross-modal association. *Proceedings of the ACM International Conference on Multimedia* 604–611.

Park, K.; Shen, C.; Hao, Z.; and Kim, J. 2011. Efficiently learning a distance metric for large margin nearest neighbor classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Rasiwasia, N.; Pereira, J. C.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. *ACM international conference on Multimedia*.

Wu, W.; Xu, J.; and Li, H. 2010. Learning similarity function between objects in heterogeneous spaces. *Microsoft Research Technical Report*.

Xing, E.; Ng, A.; Jordan, M.; and Russell, S. 2002. Distance metric learning, with application to clustering with side-information. *NIPS* 15:505–512.

Zhang, Y., and Yeung, D. 2011. Multi-task learning in heterogeneous feature spaces. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. *Advances in Neural Information Processing Systems (NIPS)*.

Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S.; Xue, G.; Yu, Y.; and Yang, Q. 2011. Heterogeneous transfer learning for image classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.